ASSESSMENT OF GAUSSIAN RADIAL BASIS FUNCTION NETWORK ON PROTEIN SECONDARY STRUCTURES

T.İbrikçi[1], M.Güler[2], M.Açıkkar[1]
[1] Department of Electrical-Electronics Engineering, Çukurova University, Adana, Turkey
[2]Department of Computer Engineering, Eastern Mediterranean University, Famagusta Mersin-10, Turkey

Abstract- **Studies of the radial basis function (RBF) network on protein secondary structures are presented. Secondary structure prediction is a useful first step in understanding how the amino acid sequence of protein determines the native state. If the secondary structure is known, it is possible to derive a comparatively small number of tertiary structures using the secondary structural element pack. A study of the Gaussian-RBF with different window sizes on the dataset developed by Qian-Sejnowski, and also a dissimilar dataset by Chandonia is given. The RBF network predicts each position in turn based on a local window of residues, by sliding this window along the length of the sequence. It is shown that the Gaussian RBF network is not an appropriate technique to be used in the prediction of secondary structure for sequence structural state.**
Keywords – **RBF Neural Networks, Protein Secondary Structure**

I.INTRODUCTION

Secondary structure prediction is an important element in understanding how the amino acid sequence of a protein determines the native state. The principles of governing protein structure are complex and not yet well understood.

Early attempts to predict the secondary structure focused on development of mapping from local windows of residues in the sequence to the structural state of the central residue in the window. Qian and Sejnowski established superior results in 1998 compared to previous studies [1][2][6][10]. They used 104-protein sets of known proteins, which were obtained from Brookheaven National Laboratory in California, USA, and extracted from several testing sets of known proteins without structural or sequence homology. They used all the sets except the test set to train the network to predict the secondary structure (helix (H), strand (S), or coil (C), and used the sigmoid equation for the transfer function. In separate study, Holley and Karplus designed a similar network for prediction of the secondary structure [3][4]. They used 48 known-protein structures that are determined by the Dictionary of Protein Secondary Structure Program (DSSP) [3]. The general rules for taking protein structures from existing databases and applying them to sequences of unknown structures currently appear to be the most practical starting point for protein structure prediction. The published results in protein literature show that neural networks have produced the most accurate secondary structure prediction with respect to the more conventional methods in the last fifteen years. The current best methods reach accuracies of about 73% with multiple homologous sequences and 70% for single sequence prediction [7][9].

II. METARIALS AND METHOD

II.1 Data sets

Two different protein data sets were used in this study. The first data set of Qian and Sejnowski classifies known structures as α-helix (H), and β-strand (S) [2]. Residues which are neither H nor S are classified as coil (C). It has 104 globular proteins that are almost identical sequences as a similar type of proteins. The database contains 21630 amino acid with 24.5% α-helix,

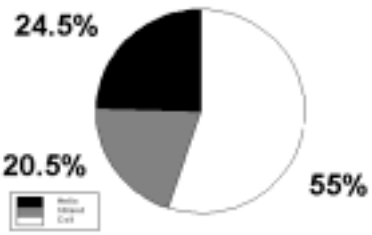

Fig. 1. The percentage of secondary structures in Qian-Sejnowski dataset

20.5% β-strand and 55 % coil. The second data set was obtained from John-Marc Chandonia, Department of Cellular and Molecular Pharmacology, University of Californina, USA. The protein used in this data set that was a set of 681 chains representative of high-resolution structures[3]. The complete database contains a total of 158428 residues and average 248 of protein with composition of 30 % α-helix (H), 22 % β-strand (E), and 48 % coil (C) that is shown in fig.2. In this study, there are trainng and test datasets for each data groups. The training set of Qian-Sejnowski contains over 87 non-homologous protein chains comprising more than 18,107 training patterns in total which equivalent to the total number residues. The testing set contains 17 protein chains with 3523 testing patterns. The training set (Chandion) covers

Report Documentation Page

Report Date	Report Type	Dates Covered (from... to)
25 Oct 2001	N/A	-

Title and Subtitle Assessment of Gaussian Radial Basis Function Network on Protein Secondary Structures	**Contract Number**
	Grant Number
	Program Element Number
Author(s)	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) Department of Electrical-Electronics Engineering, Cukurova University Adana, Turkey	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es) US Army Research, Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 09499-1500	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)

Distribution/Availability Statement
Approved for public release, distribution unlimited

Supplementary Notes
Papers from 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, October 25-28, 2001, held in Istanbul, Turkey. See also ADM001351 for entire conference on cd-rom.

Abstract

Subject Terms

Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UU

Number of Pages
3

398 protein chains with 95505 training patterns. The testing set has 283 protein chains with 62923 testing patterns.

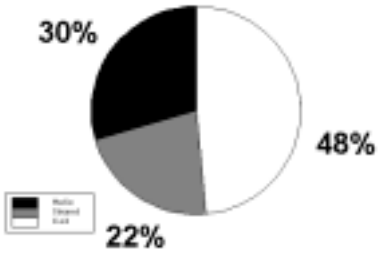

Fig.2: The percentage of secondary structures in Chandonia dataset

II.2 Structure of Network

As in most of the existing methods, the secondary structure of the I^{th} R_i is predicted from window of amino acids, R_{i-n},…..R_i,R_{i+1},….R_{i+n} where

$$\text{Winsize} = 2*n+1 \qquad (1)$$

is the window size. Each pattern is a window on to a short segment of protein chain centered on the residue (bold character below pattern sequence) to be predicted in this instance. Each central residue (bold character below target sequence) forms part of a helix, or strand or coil.

Pattern: *ENLKGF***L***VKQPEE*
Target: *CCCCCS***S***SCCCCC*

Each pattern presented to the network comprises N*M inputs for a window of size N, M is encoding which we use 20-digit encoding. The 20-digit, which is binary numbers 1 or 0, which are used to represent the identity of each protein residue in the sequence string, encodes amino acids. The string has a vector of 20 digits among which 19 have a value of 0, and one has a value of 1. Eg;
Glycine(G): 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Helix(H) : 1 0 0 (Target)

The advantage of this sparse encoding scheme is that it does not introduce an artificial correlation between the amino acids. Each amino acid and secondary structure type is given equal weight. The main disadvantage is that it entails a large number of network parameters The network architecture is a fully connected

N(20)-H-3 network,
where N is the window size (winsize), H is the number of hidden layer nodes which were 5, 10, 15, 20, 30 nodes for each network, and (20) is the encoding the digit number.

II.2 Mesurament of Accuracy

The accuracy was measured by Q_3 standard that is following equation.

$$Q_3 = \frac{H+S+C}{N} \qquad (2)$$

H, S, and C show correctly predicted helix, strand and coil, respectively, divided by the total number of N predicted residues.

II.3 Method

Radial functions are simply a class of functions. They are a kind of linear or non-linear model and any type of single-layer or multi-layer network. However, Broomhead and Lowe (1988) studied using the radial function in a single-layer radial basis function [1]. Radial basis function neural networks transform the n-dimensional inputs non-linearly to an m-dimensional space and then estimate a model using linear regression [5]. The non-linear transformation is controlled by a set of m basis functions each characterized by position or center μ in an original input space and width or radius vector μ_i i = 1,2…m. α is a shape factor.

$$h_j(x) = \exp\left(-\frac{1}{2\alpha^2}\|x-\mu_i\|^2\right) \qquad (3)$$

The basis functions are usually local that they respond most strongly to the inputs nearest to center μ_I , in the matrix determined by the radius.

$$f(x) = \sum_{j=1}^{m} w_j h_j(x) \qquad (4)$$

f(x) function is a transfer function that x shows point of data w_i are weight parameters, and h(x) is Gaussian function.

III. RESULTS AND DISCUSSION

Prediction of protein secondary structure is tested by using RBF network that traditionally has only a single hidden layer, and borrowed techniques from statistics, such as forward selection and ridge regression, as strategies for controlling model complexity.

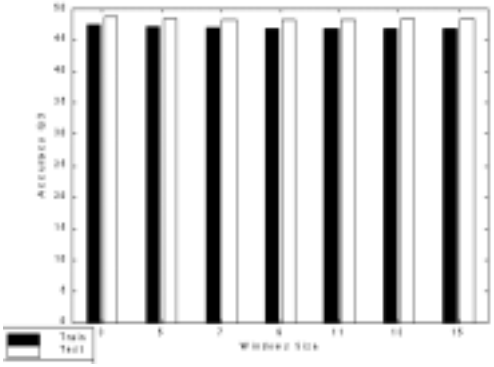

Fig. 3. Prediction Accuracy of RBF on Qian-Sejnowski Dataset

The results showed, however that the prediction values of Qian Sejnowski and Chandonia datasets are the same regardless of different used option such as window size and database size. The network could not find the residues of helix and strand, only coil was predicted with the 100% success rates. Nevertheless, the "found" values are greater that the "total" values [8].

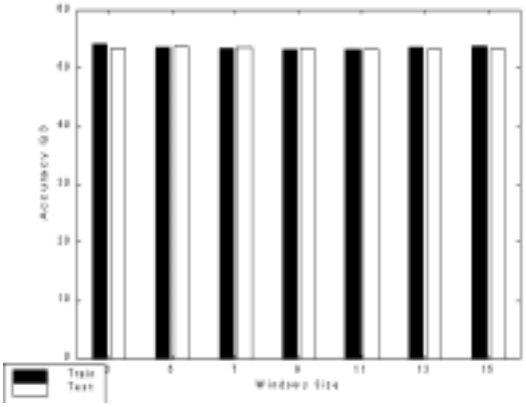

Fig. 4. Prediction Accuracy of RBF on Chandonia Dataset

The main aim of RBF is to find the center of dataset based on its density, and to locate the data, which are at the certain distance from the center. Therefore, a data grouping can be developed. If the values in the database are far from each other, classification is easy, otherwise a measure of error may appear as data can mix into each other. The distance from the center is also another criteria for causing error. If the distance is reasonably long, more data is located in the central-region. If the distance is very short, necessary data is not included in the appropriate region. Indeed, the classification is basically depending upon the actual data. Classification of the data obtained by the authors of this paper is very difficult because it consists of three residues, and the whole data is not very far from the central-region. Therefore, all the data is located in the same (one) center.

Based on the residue distribution, coil appeared more than the helix and strand: 55% coil, 20.5% helix and 24.5% strand. Since the coil is more dominant, all the residues are located under the same classification. Consequently, it shows that maximum group occurs depending on the number of data. Each data makes an individual group, but, it is not a real classification.

V. CONCLUSION

The prediction of secondary structure of proteins with radial basis functions (RBF) with different windows sizes was studied. RBF networks have already applied to many daily problems. They show good results. In this application, however, RBF networks are not a suitable means to predict the secondary structure of a protein. This is primarily because here RBF classifies all the data in the same class.

ACKNOWLEDGMENT

This research was supported by The Cukurova University Foundation MMF2000-41. We thank to Dr. Chandonia for giving permission to use his protein dataset, and to Prof. Dr. Julian Richardson who is colleague at the Department of Electrical-Electronics Engineering.

REFERENCES

[1] Broomhead, D, S., Lowe, D. "Multivariable functional interpolation and adaptive networks" *Complex System,s* Vol.2 pp:321-355, 1988.
[2] Qian, N., SejnowskI, T. J. "Predicting the Secondary Structure of Globular Proteins Using Neural Network Models", *J. Mol. Biol.* Vol:202,pp: 865-884 , 1988.
[3] ChandonIa J. M., Karplus M. "Neural networks for secondary structure and structural class prediction", *Protein Science*, Vol:4, pp: 275-285 , 1995.
[4] Holley, H., Karplus M. "Neural network for protein structure prediction", *Methods in Enzymology,* Vol: 202, pp: 204-224, 1991.
[5] Bishop, C. M. Neural Networks for Pattern Recognition, *Oxford University Press*, 1996.
[6] Matthews, B. W. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme", *Biochim. Biophys. Acta*, Vol: 405, pp: 442-451, 1975.
[7] Rost, B., Sander, C. "Prediction of protein secondary structure at better than 70% accuracy", *Journal of Molecular Biology*, Vol: 232, pp: 584-599, 1993.
[8] Ibrikci, T. Neural Network Models for Secondary Protein Structure, PhD Thesis., 2000.
[9] Solovyev, V., Salamon A. "Local secondary structure prediction using local alignments", *Journal of Molecular Biology,* pp:31-36 , 1997.
[10] Chou P.Y, Fasman G.D., "Prediction of protein", *BioChem.,* vol.13, pp.222-234, 1994